Normality of Raw Data in General Linear Models: the Most Widespread Myth in Statistics

In years of statistical consulting for ecologists and wildlife biologists, by far the most common misconception we have come across has been the one about normality in general linear models. These comprise a very large part of the statistical models used in ecology and include *t* tests, simple and multiple linear regression, polynomial regression, and analysis of variance (ANOVA) and covariance (ANCOVA). There is a widely held belief that the normality assumption pertains to the raw data rather than to the model residuals. We suspect that this error may also occur in countless published studies, whenever the normality assumption is tested *prior* to analysis. This may lead to the use of nonparametric alternatives (if there are any), when parametric tests would indeed be appropriate, or to use of transformations of raw data, which may introduce hidden assumptions such as multiplicative effects on the natural scale in the case of log-transformed data.

Our aim here is to dispel this myth. We very briefly describe relevant theory for two cases of general linear models to show that the residuals need to be normally distributed if tests requiring normality are to be used, such as *t* and *F* tests. We then give two examples demonstrating that the distribution of the response variable may be nonnormal, and yet the residuals are well behaved. We do not go into the issue of how to test normality; instead we display the distributions of response variables and residuals graphically.

A very brief theory of general linear models

We present two simple examples from among the large class of general linear models, which encompass such methods as, e.g., the *t* test, simple and multiple linear regression, polynomial regression, ANOVA, and ANCOVA. In every case, a response variable is thought to be composed of additive systematic components and one or several random components. The latter are usually assumed to be from a common normal distribution with a constant variance.

Simple linear regression

The normal error regression model for a sample of size *n* that links a response variate *Y* to one continuous explanatory variable *X* is as follows (from Neter et al. 1990:52):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where Y_i is the observed response for the *i*th unit; X_i is the value of the explanatory variable for the *i*th unit; β_0 and β_1 are parameters, i.e., unknown constants to be estimated from the data; ε_i are independent $N(0, \sigma^2)$, i.e., independent normally distributed residuals about a zero mean with constant variance σ^2; $i = 1,\ldots, n$ and indexes the units.

Normality and homoscedasticity (constant variance) of residuals is not necessary to use the least-squares or maximum-likelihood method to provide unbiased point estimates of the parameters β_0 and β_1. However, to provide significance tests or confidence intervals, the standard assumption of a normal distribution of error terms ε_i needs to be invoked (Neter et al. 1990).

One-way ANOVA

The linear additive model links a response variate *Y* to one discrete explanatory variable with *I* levels (discrete values) and can be written as (from Steel and Torrie 1980:149):

$$Y_{ij} = \mu + \tau_i + e_{ij}$$

where Y_{ij} is the observed response for the *j*th unit in group *i;* μ and τ_i are unknown constants, i.e., parameters to be estimated from the data; μ is the overall mean response and τ_i is the additive effect of level *i* ($i = 1,\ldots, I$); e_{ij} are independent random components; $j = 1,\ldots, n$ and indexes the units within each level of the explanatory variable.

Again, when significance tests or confidence intervals are desired, distributional assumptions about the random components e_{ij} need to be made. Customarily, they are assumed to be independent normally distributed with zero mean and constant variance σ^2.

Two numerical examples

Multiple linear regression

The number of fruits per stem of *Gentiana cruciata*, a rare plant of calcareous grasslands, had been measured on 810 plants and ranged from 1 to 60. The distribution of these data was clearly not normal, but right skewed (Fig. 1a). A multiple-regression model using three continuous explanatory variables (population area, number of stems per plant, and length of the longest stem) fit by the package Genstat (Payne et al. 1993) accounted for 47% of the variance in the data. It showed that the response variable was significantly and positively related to all three explanatory variables. The residuals of this model were reasonably close to a normal distribution (Fig. 1b). These data are from a larger study on a rare plant and its specialist herbivore. (For further description of the system, see Kéry et al. 2001.)

One-way ANOVA

We then generated 200 data points for each of four populations. Think of it as the mean number of seeds per fruit of *Gentiana cruciata*. Mean numbers of seeds were 100, 200, 300, and 400 in the four populations, respectively. Normally distributed noise with variance 50 was added. The distribution of these data was again far from normal (Fig. 1c). However, when the systematic popula-

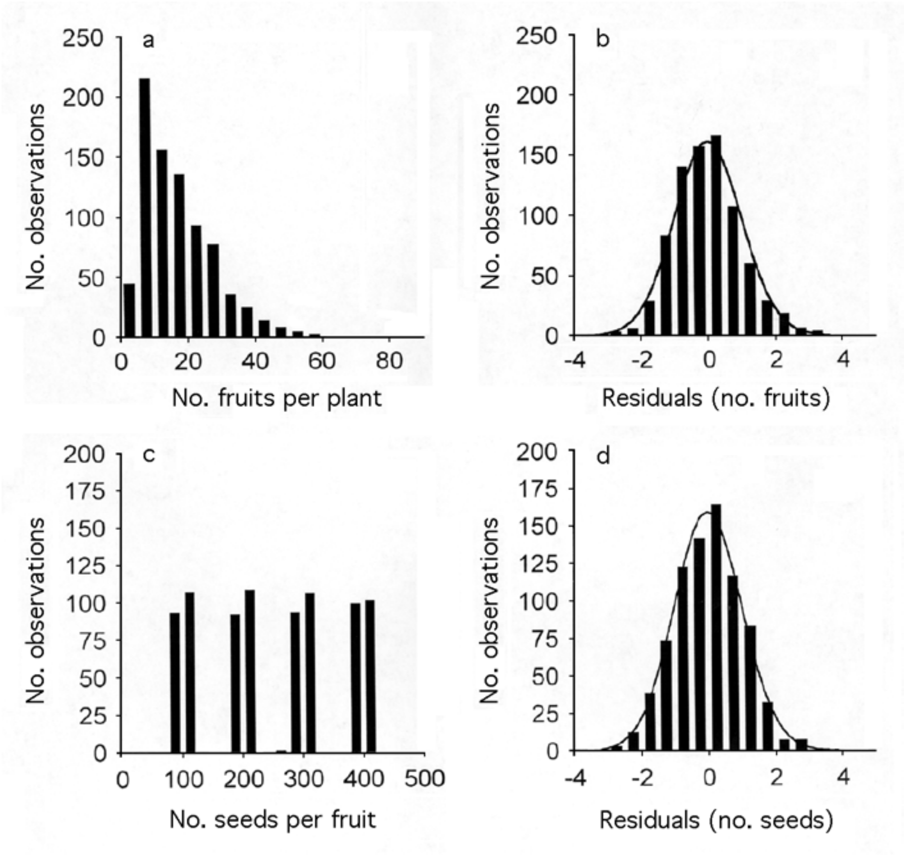

Fig. 1. Distribution of the raw data and of the residuals (a,b) for a multiple linear regression analysis, and (c,d) for a one-way analysis of variance.

tion effect was taken out by fitting one discrete explanatory variable, the resulting residuals were normally distributed (Fig. 1d), as would be expected in this simulated case.

Conclusions

There is a very widespread misconception that, in general linear models, the raw data instead of the residuals of a model have to be normally distributed to permit construction of confidence intervals and significance statistics. Here we state this to be false and give two examples that show raw data may have some other distribution, and yet the residuals of a linear model turn out to be reasonably close to a normal distribution. Such examples abound, and we think that only in a minority of cases are the raw data already clustered symmetrically around a single mode. Residual analysis for linear models is easily conducted in the two statistical packages that we are familiar with, Genstat (Payne et al. 1993) and SAS (SAS 2001). Residuals are easily stored in each analysis and then examined visually, e.g., by histograms or plots, or by formal statistical tests for normality. There is a large literature on model checking, also called model criticism, in general linear models (e.g., Cook and Weisberg [1982], Atkinson [1985], and also general texts on regression such as Draper and Smith [1981], or Neter et al. [1990]). Model criticism is an important part of any statistical modeling. In summary, we hope that this note is a contribution toward better statistical practice by doing away with the myth of normality of the raw data in general linear models.

Literature cited

Atkinson, A. C. 1985. Plots, transformations and regression. An introduction to graphical methods of diagnostic regression analysis. Oxford University Press, Oxford, UK.

Cook, R. D., and S. Weisberg. 1982. Residuals and influence in regression. Chapman and Hall, London, UK.

Draper, N. R., and H. Smith. 1981. Applied regression analysis. Wiley, New York, New York, USA.

Kéry, M., D. Matthies, and M. Fischer. 2001. The effect of plant population size on the interactions between the rare *Gentiana cruciata* and its specialized herbivore *Maculinea rebeli*. Journal of Ecology **89**:418–427.

Neter, J., W. Wasserman, and M. H. Kutner. 1990. Applied linear statistical models. Third edition. Irwin, Burr Ridge, Illinois, USA.

Payne, R. W., P. W. Lane, P. G. N. Digby, S. A. Harding, P. K. Leech, G. W. Morgan, A. D. Todd, R. Thompson, G. Tunnicliffe Wilson, S. J. Welham, and R. P. White.

1993. Genstat 5, Release 3. Reference manual. Clarendon Press, Oxford, UK.

SAS. 2001. Statistical analysis system. Version 8.02. SAS Institute, Cary, North Carolina, USA.

Steel, R. G. D., and J. H. Torrie. 1980. Principles and procedures in statistics. A biometrical approach. Second edition. McGraw-Hill, Auckland, New Zealand.

Marc Kéry and Jeff S. Hatfield
USGS Patuxent Wildlife Research Center
11510 American Holly Drive
Laurel, MD 20708
(301) 497-5632
Fax: (301) 497-5666,
E-mail: mkery@usgs.gov